

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 10-288996

(43)Date of publication of application : 27.10.1998

(51)Int.Cl.

G10L 3/00
G10L 3/00

(21)Application number : 10-097547

(71)Applicant : NOKIA MOBILE PHONES LTD

(22)Date of filing : 09.04.1998

(72)Inventor : LAURILA KARI
VIKKI OLLI

(30)Priority

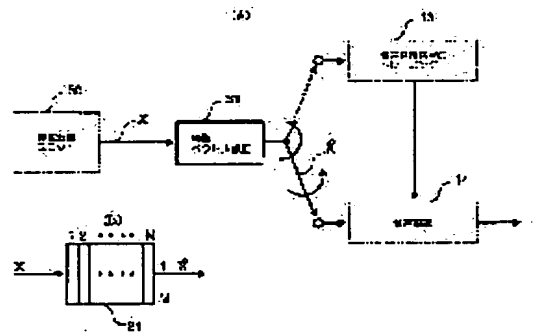
Priority number : 97 971521 Priority date : 11.04.1997 Priority country : FI

(54) SPEECH RECOGNITION METHOD AND SPEECH SIGNAL PROCESSOR

(57)Abstract:

PROBLEM TO BE SOLVED: To correct a feature vector to be decided at a speech recognition time for correcting the effect of a noise by defining the mean value and the standard deviation of the characteristic vector, using these parameters and normalizing the characteristic vector.

SOLUTION: The mean value and the standard deviation of the characteristic vector are defined, and the feature vector is normalized using these parameters. In such a case, the feature vector is normalized using a sliding normalization buffer preferably. In a speech recognition device, a front end 30 forms the feature vector X_i ($i=1-M$) at a prescribed interval. The feature vector X is stored in a normalization buffer 31, and thus, the mean value and the standard deviation related to respective feature vector components X_i are calculated. Thereafter, the feature vector component X_i to be recognized is normalized using a calculated normalization coefficient in the block 31. In such a case, the normalization buffer 31 that e.g. a length N is fixed is slid on the feature vector X .



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision
of rejection]

[Date of requesting appeal against examiner's
decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平10-288996

(43) 公開日 平成10年(1998)10月27日

(51) Int.Cl.^a

G 1 0 L 3/00

識別記号

5 3 5

5 2 1

F I

G 1 0 L 3/00

5 3 5

5 2 1 F

審査請求 未請求 請求項の数 6 O L (全 9 頁)

(21) 出願番号 特願平10-97547

(22) 出願日 平成10年(1998)4月9日

(31) 優先権主張番号 9 7 1 5 2 1

(32) 優先日 1997年4月11日

(33) 優先権主張国 フィンランド (F I)

(71) 出願人 590005612

ノキア モービル フォーンズ リミティ
ド

フィンランド国, エフアイエヌ-02150

エスボー, ケイララーデンティエ 4

(72) 発明者 カリ ローリラ

フィンランド国, エフアイエヌ-33720

タンペレ, インシンオーリンカトゥ 64
アー 14

(72) 発明者 オリ ビッキー

フィンランド国, エフアイエヌ-33100

タンペレ, ビンニンカトゥ 21 アー 7

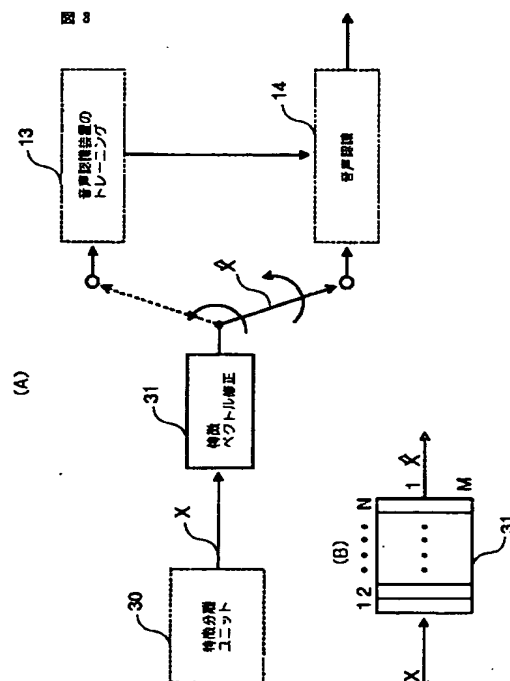
(74) 代理人 弁理士 石田 敬 (外4名)

(54) 【発明の名称】 音声認識方法及び音声信号処理装置

(57) 【要約】

【課題】 改良された音声認識方法及び音声信号処理装置を提供する。

【解決手段】 音声認識装置の分析ユニットで作られた特徴ベクトルが雑音の影響を補償するために修正される。本発明では、スライディング正規化バッファ(31)を使って特徴ベクトルを正規化する。本発明の方法により、音声認識装置のトレーニング段階が実際の音声認識段階での騒音環境とは異なる騒音環境で実行された場合に、音声認識装置の性能が向上する。



(2)

特開平10-288996

【特許請求の範囲】

【請求項1】 認識可能な音声信号を時間上で特定の長さの連続するフレームに分割し、フレームあたりに少なくとも1つの、該音声信号を説明するパラメータを作るために各音声フレームを分析し、特定のフレームに関連する前記パラメータを記憶し、前記パラメータを修正し、その修正されたパラメータを使って音声認識を実行する音声認識方法であって、連続するパラメータのうちの一部だけを定期的に記憶し、前記の修正されたパラメータを作るために定期的に記憶されるパラメータに基づいて少なくとも1つのパラメータを修正するようになっていることを特徴とする音声認識方法。

【請求項2】 N個の連続するパラメータに基づいて、次の各量すなわち平均値及び標準偏差のうちの1つを前記修正のために確定するようになっており、Nは整数であることを特徴とする請求項1に記載の方法。

【請求項3】 パラメータの前記修正は、前記各量の1つに関連する正規化から成ることを特徴とする請求項2に記載の方法。

【請求項4】 音声信号を時間上で分割して連続するフレームとするための手段(21)と、音声フレームを分析して該音声信号を説明する少なくとも1つのパラメータを作るための手段(11, 30)と、該パラメータを記憶するための記憶手段(31)と、前記パラメータを修正して修正済みパラメータを作るための手段(31)と、その修正済みパラメータを使って音声認識するための手段(14)とから成る音声信号処理装置であって、前記記憶手段(31)は前記の連続するパラメータのうちの一部だけを定期的に記憶するようになっており、該パラメータを修正するための前記手段(31)は、前記修正済みパラメータを作るために該記憶手段(31)に定期的に記憶されたパラメータに基づいて該音声信号を説明する該パラメータを修正するようになっていることを特徴とする音声信号処理装置。

【請求項5】 前記記憶手段(31)は一定の長さのバッファ(31)から成ることを特徴とする請求項4に記載の装置。

【請求項6】 前記記憶手段(31)は長さが変化し得るバッファ(31)から成ることを特徴とする請求項4に記載の装置。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は音声認識に関し、特に音声認識において決定されるべき各特徴ベクトル(feature vector)を修正する方法に関する。本発明は、音声認識を改良するために本発明のこの方法を使用する装置にも関する。

【0002】

【従来の技術】本発明は自動的音声認識に関し、特に、ヒドゥン・マルコフモデル(Hidden Markov Models (HM

M))に基づく音声認識に関する。HMMに基づく音声認識は、認識可能な単語の統計的モデルに基づいている。認識段階においては、発音された単語についてマルコフチェーンに基づいて観測結果及び状態遷移が計算されて、音声認識装置のトレーニング段階で記憶された、その発音された単語に対応するモデルが確率に基づいて決定される。例えば、ヒドゥン・マルコフモデルに基づく音声認識方法は下記の参考文献において解説されている：“1989年2月のIEEE会報第77巻第2号の中のL. ラビナーの”音声認識におけるヒドゥン・マルコフモデルと選択されたアプリケーションについての指導”(“L. Rabiner, “A tutorial on Hidden Markov Models and selected applications in speech recognition”, Proceedings of the IEEE, Vol. 77, No.2, February 1989.

【0003】

【発明が解決しようとする課題】現在の音声認識装置の問題は、騒々しい環境の中では認識精度が著しく低下することである。また、特に音声認識装置が動作するときの騒音条件が該音声認識装置のトレーニング段階での騒音条件と異なる場合には該音声認識装置の性能は低下する。音声認識装置が使用されることのある全ての騒音環境の影響を考慮に入れることは不可能であるので、このことは実際問題としては音声信号認識システムにおいて最も解決しにくい問題の1つである。音声認識装置を利用する装置のユーザにとっての正規の立場は、音声認識装置のトレーニングは通常は殆ど騒音のない環境で行われるけれども、その音声認識装置が例えば車内などの動作環境で使われるときには周囲の交通やその車自体から生じる暗騒音はトレーニング段階での殆ど静穏な暗騒音レベルとは著しく違っているということである。

【0004】音声認識装置の性能が使用されるマイクロホンに左右されることも現在の音声認識装置の問題である。特に音声認識装置のトレーニング段階で使われるマイクロホンが実際の音声認識段階で使われるマイクロホンとは違っている場合には、その音声認識装置の性能は著しく低下する。

【0005】特徴ベクトルを計算する際の雑音の影響を消去するために幾つかの方法が開発されている。しかし、それらの方法を利用する音声認識装置は決まったコンピュータ/ワークステーションのアプリケーションでのみ利用可能であり、それらの音声認識装置では音声はオフラインで認識される。それらの方法では、認識されるべき音声はコンピュータのメモリーに記憶されるのが普通である。通常、記憶される音声信号の長さは数秒である。その後、特徴ベクトルは、計算時に、ファイル全体の内容から確定される各パラメータを使って修正される。記憶される音声信号の長さの故に、その種の方法を実時間音声認識に適用することはできない。

【0006】また、正規化方法も設けられており、その

(3)

特開平10-288996

方法では音声及び雑音の両方が自分の正規化係数を持っていて、それらは音声活性検出器(VAD)を用いて適応的に更新される。適応的更新(adaptive updating)であるために、各正規化係数の更新には遅延が伴うので、正規化プロセスは実用上十分な速さでは実行されない。また、この方法もVADを必要とするけれども、その動作は、信号対雑音比(SNR)の値が低い音声認識アプリケーションではしばしば余りに不正確である。この方法も、前記の遅延の故に実時間要件を満たさない。

【0007】

【課題を解決するための手段】上記の問題を解決する音声認識方法及び装置が発明されており、その方法及び装置により音声認識時に決定される特徴ベクトルは雑音の影響を補償するために修正される。特徴ベクトルの修正は、特徴ベクトルの平均値と標準偏差とを定義し、それらのパラメータを使って特徴ベクトルを正規化することによって実行される。本発明の好ましい実施例では、スライディング正規化バッファ(sliding normalisation buffer)を使って特徴ベクトルを正規化する。本発明では、特徴ベクトルの正規化パラメータの更新は殆ど遅延無しで実行され、実際の正規化プロセスにおける遅延は十分に小さいので実時間音声認識アプリケーションを実現することができる。

【0008】また、本発明の方法によれば、音声認識装置の性能を、使用するマイクロホンに左右されにくくすることができる。本発明によれば、音声認識装置の実験段階と認識段階とで異なるマイクロホンが使われる場合にも、同じマイクロホンがトレーニング段階及び認識段階で使われる場合と殆ど同じ程度に、音声認識装置の高い性能が達成される。

【0009】本発明は、請求項1及び4の特徴付け部分に記載されている事項を特徴とする。

【0010】

【発明の実施の形態】図1は、本発明に適用できる公知の音声認識装置の構造を示すブロック図である。通常、音声認識装置の動作は、図1に示されているように、主要な2種類の活動、即ち実際の音声認識段階10-12、14-15と音声トレーニング段階13とに分けられる。音声認識装置はマイクロホンから入力として音声信号 $s(n)$ を受け取り、この信号は、例えば8kHzのサンプリング周波数及び1サンプルあたり12ビットの分解能を使用するA/D変換器10によってデジタル形に変換される。通常、音声認識装置はいわゆるフロント・エンド11を有し、ここで音声信号が分析されて特徴ベクトル12がモデル化される。特徴ベクトルは特定の期間中の該音声信号を描写するものである。特徴ベクトルは、例えば10ms間隔で確定される。特徴ベクトルを、数種類の手法でモデル化することができる。例えば、特徴ベクトルをモデル化するための数種類の手法が

下記の参考文献で解説されている：1993年9月のIEEE会報第81巻、第9号、pp. 1215-1247、に掲載されているJ. バイコーンの”音声認識における信号モデル化手法”(J. Picone, "Signal modelling techniques in speech recognition", IEEE Proceedings, Vol. 81, No. 9, pp. 1215-1247, September 1993. 本発明において使用される特徴ベクトルは、いわゆるメル周波数セブストラル係数(Mel-Frequency Cepstral Coefficients (MFCC))を確定することによりモデル化される。トレーニング段階で、音声認識装置により使用される単語について音声認識装置のトレーニング・ブロック13において、特徴ベクトルによってモデルが作成される。モデル・トレーニング13aにおいて、認識可能な単語についてモデルが決定される。トレーニング段階において、モデル化されるべき単語の復唱(repetition)を利用することができる。モデルはメモリー13bに記憶される。音声認識時に、特徴ベクトルは現実の認識装置14に送られ、この装置は、ブロック15aにおいて、トレーニング段階時に構成されたモデルと認識可能な音声から構成されるべき特徴ベクトルとを比較して、認識結果についての判定をブロック15bで行う。認識結果15は、音声認識装置を使用する人により発音された単語に最もよく対応する、音声認識装置のメモリーに記憶されている単語を表示する。

【0011】図2は、本発明に適用できるフロント・エンド11の公知の分析ブロックの構造を示している。通常、フロント・エンド11は、音声認識に関連する周波数を強調するためのプリエンファシス・フィルター20を有する。通常、プリエンファシス・フィルター20は、例えば、 $H(z) = 1 - 0.95z^{-1}$ のレスポンスを有する1次FIRフィルターなどの高域通過フィルターである。次に、ブロック21において、フィルタリングされた信号からNサンプルの長さの各フレームが形成される。例えば、 $N = 240$ のサンプル長を使って、8kHzのサンプリング周波数で30msのフレーム構造が作られる。通常、連続するフレーム同士がS個の連続するサンプル(例えば10ms)の程度に重なり合ういわゆるオーバーラップ手法を使って各音声フレームを形成することもできる。ブロック23において音声信号について高速フーリエ変換(FFT)周波数表示をモデル化する前に、例えば、ブロック22においてハミングウィンドウ(Hamming window)などを使ってスペクトル推定値の精度を向上させるためにいわゆるウィンドウイング(windowing)を実行することもできる。次に、信号のFFT表示をメル・ウィンドウイング・ブロック(Mel windowing block)24においてメル周波数表示に変換する。メル周波数表示への変換は、それ自体としては当業者に知られている。メル周波数表示への変換は参考原典”IEEE会報第81巻、第9号に掲載されているJ. バイコー

(4)

特開平10-288996

ンの”音声認識における信号モデル化手法(J. Picone, "Signal modelling techniques in speech recognition")”で解説されている。この周波数変換で、いろいろな周波数に対する耳の非線形の感度を考慮に入れる。通常、使用される周波数帯域の数(k)は $k=24$ であってよい。実際の特徴ベクトル12, 即ちいわゆるセブストラル係数 $c(i)$ は、ブロック25で形成された26個の対数メル値に対していわゆる離散余弦変換(discrete cosine transformation)(DCT)を実行することによって得られる。この離散余弦変換に例えば次数 $J=24$ を使用することができる。通常、DCT係数 $c(i)$ (i は余弦項のインデックスである)のうちの半分だけが使われる。通常、実際の特徴ベクトルは、いわゆる第1段及び第2段の差信号 $dc(i)$ 及び $ddc(i)$ を計算することによって音声の変動過程(ダイナミクス)に関する情報も包

含する。ブロック27において $dc(i) = c(i) - c(i-1)$ 及び $ddc(i) = dc(i) - dc(i-1)$ を推定することにより、離散余弦変換ブロックの連続する出力ベクトルからこれらの差信号を決定することができる。これらの26個の追加のパラメータが考慮される場合には、特徴ベクトルの長さは例えば $13+26=39$ パラメータとなる。

【0012】図3(A)及び(B)は本発明の第1実施例の音声認識装置の構造を示す。フロント・エンド30は10ms間隔で出力信号として特徴ベクトル X_i , $i=1 \dots M$ (例えば $M=39$)を作成する。特徴ベクトルは正規化バッファ31に記憶され、これにより各特徴ベクトル成分 X_i , $i=1 \dots M$, についての平均値 μ_i 及び標準偏差 σ_i が次のように計算される:

【数1】

$$\mu_i = \frac{1}{N} \sum_{t=1}^N x_{t,i}, i=1, \dots, M \quad 1$$

【数2】

$$\sigma_i = \sqrt{\frac{1}{N} \sum_{t=1}^N (x_{t,i} - \mu_i)^2}, i=1, \dots, M \quad 2$$

式(1)及び(2)において、 N は正規化バッファ(normalisation buffer)の長さであり、 M は特徴ベクトル(feature vector)の長さである。この後、ブロック31において、計算された正規化係数 μ_i , σ_i を使って、認識される

$$\hat{x}_i = \frac{x_i - \mu_i}{\sigma_i}, i=1, \dots, M \quad 3$$

【0013】段落番号【0013】から【0014】までに現れる(イ)は、下記表1に示す記号を表す。

【表1】

(イ)	\hat{x}
-----	-----------

正規化された特徴ベクトル(イ)はそれが音声認識装置のトレーニング段階であるのか実際の音声認識段階であるのかにより、音声認識ユニット14又はトレーニング・ブロック13に入力信号として送られる。本発明の第1実施例の方法では、長さ(N)が固定されている正規化バッファが使われ、このバッファは特徴ベクトル上をスライドさせられる。このスライディング正規化バッファがあるので、この方法を実時間音声認識システムで実行することもできる。正規化バッファ31は $N * M$ サンプルのサイズのバッファであり、通常は、デジタル信号処理装置(DSP)の内部メモリ構造又は外付けメモリを使って該DSPをプログラミングすることによって音声認識ユニットと関連させてこのバッ

べき特徴ベクトルの成分 X_i が正規化される。正規化されるべき特徴ベクトル X は、図4に示されているように正規化バッファ31の中央に置かれる。

【数3】

ァーを実現することができる。本発明の実施例の解決法においては、正規化バッファは100の特徴ベクトルの長さを持っている。一度に正規化されるべき特徴ベクトルは正規化バッファ31の中央に置かれる。正規化されるべき特徴ベクトルは正規化バッファの中央に置かれるので、音声認識には正規化バッファの長さである N の遅延が生じる。本例の各パラメータを使うときには、遅延は $100 * 10ms = 1$ 秒である。しかし、次に説明するように音声認識の始めに該バッファの長さの一部分だけを使うことによって、この遅延を半分にすることができる。

【0014】図5及び図6は、フローチャートの形で、本発明の方法の作用を示している。音声認識の始めに、正規化バッファの全長の半分 $N/2$ が使用されるまで該正規化バッファは充填される(ブロック100-102)。その後、平均値及び標準偏差の各ベクトル μ_i , σ_i が計算され(ブロック103)、はじめの $N/2$ 個の特徴ベクトルを使って第1特徴ベクトルが正規

(5)

特開平10-288996

化される。ブロック15b(図1)で公知の手法に従ってビタビ復号(Viterbi decoding)によりこの正規化された特徴ベクトル(イ)に対して実際の音声認識プロセスが実行される。次に、新しい特徴ベクトルが緩衝記憶され(ブロック104)、記憶されている($N/2+1$)個の特徴ベクトルを使って新しい正規化係数が計算され、第2の特徴ベクトルが正規化されて、それに対して認識が実行される(ブロック103)。これに対応するプロセスが正規化バッファが満杯になるまで続けられる。このときフローチャートにおいてブロック105からブロック106への移行が行われる。このことは、始めの $N/2$ 個の特徴ベクトルが認識され終わっていて、正規化されるべき特徴ベクトルが正規化バッファの中央に位置していることを意味する。このとき該バッファはFIFO原理(先入れ先出し)に従ってスライドされて、新しい特徴ベクトルが計算され認識され終わったならば(ブロック107)、最も古い特徴ベクトルが正規化バッファから除去される(ブロック106)。認識段階の終わりに(ブロック107)、正規化バッファに記憶されている値を使って正規化係数が計算される。これらの正規化係数が最後の $N/2$ 個の特徴ベクトルの認識と関連して使用される。平均値及び標準偏差は、正規化されていない特徴ベクトルを使って計算される。 N 個の特徴ベクトルの全てに対して音声認識が実行され終わると(ブロック108)、音声認識装置は認識可能な単語の結果をモデル化する(ブロック109)。

【0015】本発明の第2の実施例では、正規化バッファの長さは音声認識中に変化することがある。音声認識開始時には長さが比較的に短い(例えば $N=45$)バッファを使うことができ、例えば各フレーム(30ms)について音声認識が進むに連れて、緩衝記憶されるべき信号の長さを大きくしてゆくことができる。この様に、本発明の第1実施例に対する例外として、正規化されるべき特徴ベクトルはバッファの中央の特徴ベクトルではなくてバッファに最初にロードされた特徴ベクトルであってもよく、そのときのバッファの内容の全部を正規化係数の計算に利用することができる。この応用例では、遅延の長さは N であり、 N は音声認識の始めでのセグメントの長さである(例えば、 $N=45$)。

【0016】本発明の1実施例では、特徴ベクトルの成

分の全てが正規化されるのではなくて、特徴ベクトルの成分のうちの一部に対してだけ正規化が実行される。例えば、人の聴覚作用/音声認識に関して最も重要な成分だけに対して正規化を実行してもよい。また、本発明の変形例では、平均値又は標準偏差と関連させて特徴ベクトルに対して正規化を実行するだけでもよい。より一般的に、特徴ベクトルの修正を如何なる統計量に関連させて実行してもよい。

【0017】図7は移動局の構造を示しており、この移動局には、本発明を利用する音声認識装置66が設けられている。この移動局は、該装置に特有の例えばマイクロホン61、キーボード62、ディスプレイ63、スピーカー64及び制御ブロック65などの部分からなっており、この制御ブロックは該移動局の動作を制御する。また、この図は、移動局に特有の送信ブロック67及び受信ブロック68も示している。制御ブロック65は、該移動局と関連している音声認識装置66の動作も制御する。この音声認識装置がそのトレーニング段階又は実際の音声認識プロセス時に活性化されているとき、ユーザーが与えるオーディオコマンドが制御ブロックによって制御されてマイクロホン61から音声認識装置66に送られる。オーディオコマンドは別のHF(hands free(ハンドフリー))マイクロホンを通して送られてもよい。通常、音声認識装置はDSPによって実現され、その動作に必要なROM/RAMメモリ回路を有する。

【0018】表2は本発明の方法での音声認識装置の性能を他の騒音補償方法と比べて示している。本発明は、正規化されていないメル周波数セブストラル係数又はPMC(Parallel Model Combination(並列モデル結合))法の使用と比較されている。試験は、騒音の少ない環境でモデル化されたヒドン・マルコフ・モデルを使って実行された。音声認識時には、必要な信号対雑音比を達成するために、認識されるべき単語に雑音信号が加えられた。"クリーン"モードは、音声認識装置のトレーニングと実際の音声認識プロセスとがともに騒音の少ない環境で行われた事態に相当する。試験結果は、本発明の音声認識装置が特に騒々しい環境で認識装置の信頼性を向上させることを証明している。また、本発明の音声認識装置は、計算に関しては本発明の方法よりはるかに複雑なPMC法より良好に機能することが分かる。

【表2】

環境(SNR)	MPCC-係数	PMC	正規化された特徴ベクトル
クリーン	96.5%	96.6%	97.5%
5dB	95.0%	95.3%	96.1%
0dB	93.7%	94.9%	95.9%
-5dB	89.3%	93.0%	95.3%
-10dB	73.8%	84.6%	94.3%

【0019】本明細書では本発明を具体例により説明し

ている。例えば、上の解説では、HMMに基づく音声認

(6)

特開平10-288996

識装置で本発明を解説している。しかし、本発明は他の手法に基づく音声認識装置に用いるのにも適している。例えば、ニューラル・ネットワークを利用する音声認識装置に本発明を適用することができる。本発明は上記の実施例の詳細に限定されるものではなく、本発明の特徴から逸脱せずに本発明を他の形でも実施し得ることは当業者にとっては明らかなことである。上記実施例は、限定をするものではなくて実例であると解されるべきものである。従って、本発明を実施し使用する可能性は特許請求の範囲の各請求項のみによって限定される。従って、各請求項により確定される、均等実施態様を含む、本発明のいろいろな実施態様も本発明の範囲内に属する。

【図面の簡単な説明】

【図1】従来技術の音声認識装置の構造を示すブロック図である。

【図2】従来技術による分析ブロックの構造を示すブ

ック図である。

【図3】(A)及び(B)は、本発明の音声認識装置の構造を示す図である。

【図4】本発明による正規化バッファの使用を示す図である。

【図5】本発明による方法の作用を示すフローチャート(その1)である。

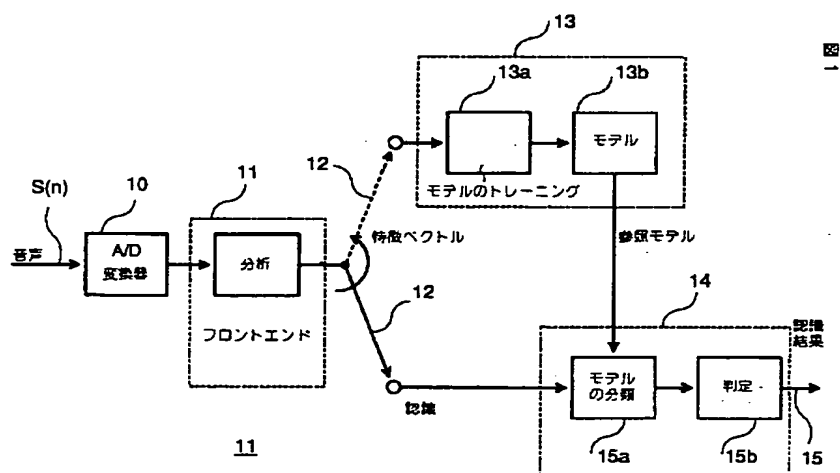
【図6】本発明による方法の作用を示すフローチャート(その2)である。

【図7】本発明の移動局の構造を示す図である。

【符号の説明】

- 11、30…フロント・エンド
- 13…音声認識装置のトレーニングブロック
- 14…現実の音声認識装置
- 20…プリアンファシス・フィルター
- 21…フレーム形成ブロック
- 31…正規化バッファ

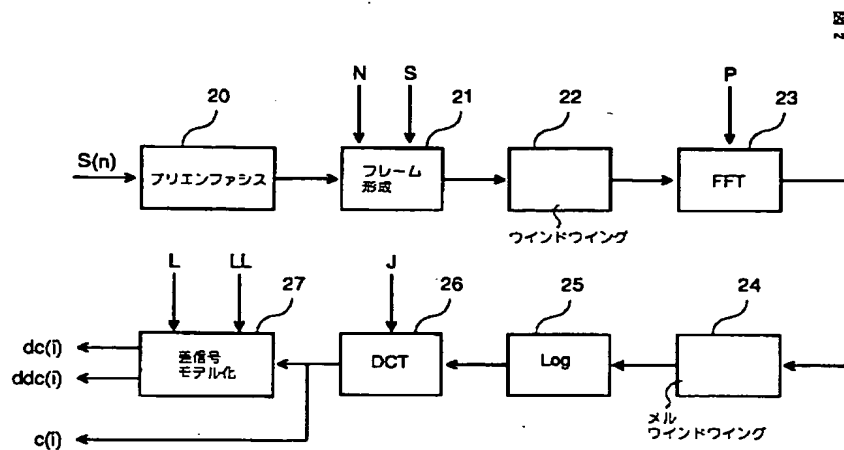
【図1】



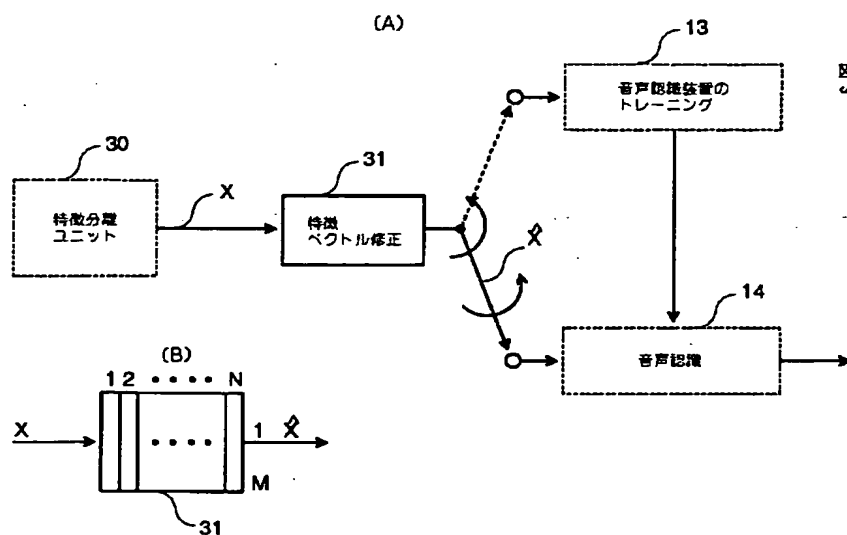
(7)

特開平10-288996

【図2】



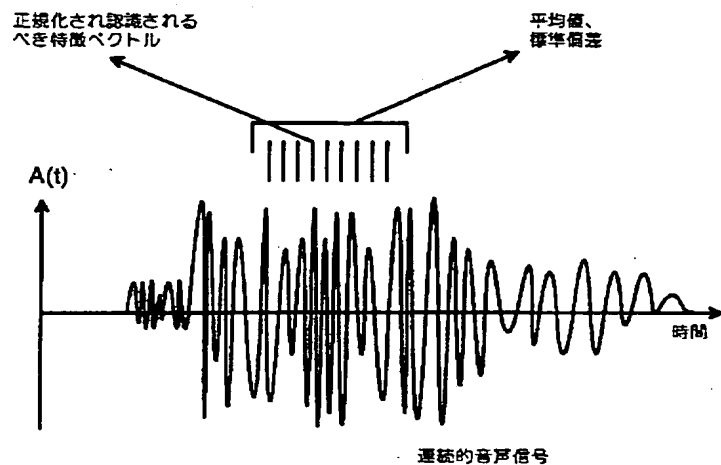
【図3】



(8)

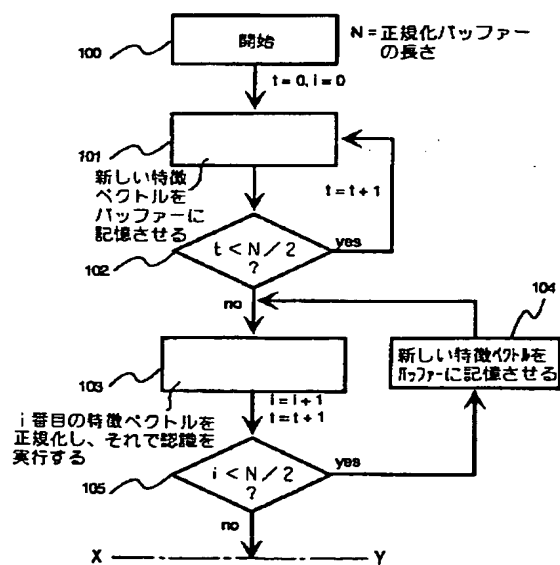
特開平10-288996

【図4】



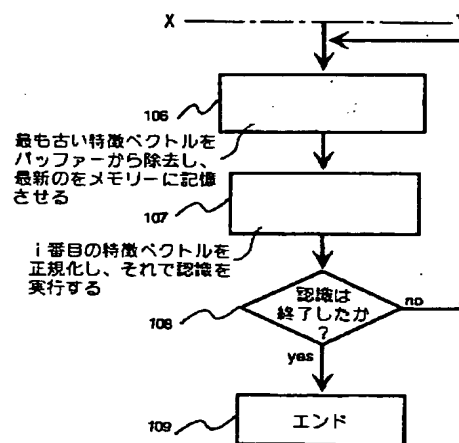
【図5】

図 5



【図6】

図 6



(9)

特開平10-288996

【図7】

